International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

Prediction of Heart Disease Using Machine Learning Algorithms

[*]Yegamati Akhila[1], R. Usha Rani[2]

[1]M.Tech.-A.I. student, Department of CSE, CVR College of Engineering, Vastunagar, Mangalpally, Ibrahimpatnam, Telangana, India

[2]Professor, Department of CSE (AI&ML), CVR College of Engineering, Vastunagar, Mangalpally, Ibrahimpatnam, Telangana, India

Article Info

Publication Issue :
Volume 8, Issue 5
September-October-2022

Page Number : 273-282

Article History
Accepted: 02 Oct 2022
Published: 29 Oct 2022

ABSTRACT

Because of modern living habits, heart disease has become a leading cause of death worldwide. Precise diagnosis and early treatment are becoming increasingly important as time goes by. We have used machine learning models and a voting classifier with data from a small sample of people to make predictions about this potentially fatal disease. This data includes the participants' medical histories and demographic information, such as whether or not any members of their immediate families have had heart problems in the past. This research presents a preliminary investigation and analysis by employing a variety of machine learning methods, including KNN, Logistic Nave Bayes, Support Vector Machine, Logistic Regression, and ensemble algorithms ADABoost and XGBoost as voting classifiers. The goal of this study is to determine if a person will develop heart disease. In comparison to other methods, the prediction accuracy of the voting classifier is 98.3%, showing that it performs quite well.

Keywords : Support Vector Machine, Logistic Regression, Naïve Bayes, KNN, ADAboost, XGBoost, and Machine Learning.

I. INTRODUCTION

The process of determining how likely an individual is at risk of getting a particular disease is an important one in the medical field and may be used to great effect in the prevention of that condition. As an added bonus, early diagnosis can be used to maximise the therapeutic benefits of treatment. Machine learning may be a better option for achieving high accuracy for predicting heart disease because this flexible tool uses feature vectors and various data types under various conditions. Heart disease is hard to predict because it depends on a lot of different factors and is very technical[1]. Risk assessment for diseases is an important tool for disease prevention in the medical sector. Increased pharmaceutical effectiveness can also be achieved by early identification. Numerous studies have been conducted over the years to determine the causes and discover novel treatments for cardiac arrest. Disease diagnosis is the primary focus of the healthcare industry[2] The ability to detect disease in its earliest stages has the potential to save many lives. If machine learning

Copyright: © the author(s), publisher and licensee Technoscience Academy. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited

identification algorithms could help doctors diagnose diseases more quickly and accurately, they might significantly advance medical science. Men have a higher incidence of cardiac problems than women do. Predicting and diagnosing heart disease in this study involves using factors like heart rate, blood pressure, gender, diabetes, age, and so on. Many different factors combine to make heart disease difficult to diagnose in advance. Cholesterol, smoking, weight, hypertension, hereditar y predisposition, and stress at work are just some of th e risk factors for cardiovascular disease. Predicting car diac issues requires an essential and precise role for m achine learning algorithms.

The main contribution of the paper is to use focus on a different machine classifier and an extreme learning machine as feature extractor in order to predict CHD in a large-scale, using the data of electronic medical records.

The remaining paper is organized as follows: Section 2 presents the related work. Section 3 discusses the system architecture, Section 4 discusses implementation and experiment details, Section 5 provide experimental results and analysis, and Section 6 concludes the paper.

II. LITERATURE REVIEW

Nowadays, more people die from heart disease than for any other single reason. Although distinguishing between evidence and cardiovascular illness is a crucial and intricate activity that calls for a cautious and skilled application, the prospect of automating this process is quite alluring. People are not born with expert-level abilities. Despite common belief, not every famous person is a gifted professional, and many areas lack easy access to those with the necessary authority and gift. With the use of the mechanical framework for therapeutic analysis, inhalation evaluation has been made more accurate and expenses have been cut.

The approach created by Purushottam et al. [3] in this work allows us to efficiently identify theories to measure the amount of patient risk in relation to a specified health parameter. The primary value of this research is to help general practitioners assess their patients' potential for developing cardiovascular disease. Real rules, shortened rules, duplication rules, eligibility rules, controlled rules, and polishing all take a back seat to the guidelines provided by their suggested system. The findings of an evaluation of the system's capacity to more properly identify the risk of coronary heart disease indicate that its adoption has been highly successful. As a result of their efforts, the authors have developed a reliable method for using data mining to forecast the occurrence of heart disease. To put it another way, the system aids doctors in making sound decisions within a certain framework.

Analysis of the risk of incident Cardiovascular Disease (CVD) based on a variety of data quality and completion rates in Electronic Health Records and Generally Collected Data shows the need for a thorough eye exam. A longitudinal cohort study served as the basis for the research design. As it relates to hospital admissions data, 392 common operations are included (for a total of 3.6 million patients). These techniques, which Li Yan [4] (2019) uses, include Sez coherence measures, which evaluate the externality of each practice, and were utilised to evaluate the data quality changes that had taken place. Statistical degradation models (linear predictor) were used to look at both the effect of overall risk variables and the difference in how accurate QRISK3 estimates were.

There are substantial variations in CSD event procedures that are not captured by QRISK3. QRISK3 estimates a 10% risk for women even with the mildest statistical disruption, rising to 7.1% to 9.0% if practise variability is factored into statistical impairment models, and hovering in the range of 10.9% to 16.4% across the majority of the quintiles. Differences in statistical degradation can be contrasted with variations in data quality (through intelligent metrics) and data coverage. Variations in data quality or the

influence of risk factors alone cannot account for the observed significant variation in the prevalence of cardiovascular disease among different populations. QRISK3 risk prediction should be supplemented by clinical judgement and other risk factor information.

A description of machine learning techniques for the diagnosis of diabetes and cardiovascular disease is provided by the work of Alic et al. [5]. (CVD). To do this, the authors have employed a hybrid approach, including both artificial neural networks (ARNs) and Bayesian networks (BNs). Certain papers published between 2008 and 2017 were subjected to a cross-article comparison. By employing a standard ANN-style learning technique called Levenberg-Marquard, a multilayer neural network may take direct action. But the most commonly used form of BN network is the naive Bayesian network, which achieves retrospective 99.51% and 97.92% accuracy in diabetes and CVD classification, respectively. Furthermore, the mean precision calculation of the experimental networks revealed improved results when utilising RNA, suggesting a greater likelihood of more accurate findings when using RNA for the categorization of diabetes and/or CVD.

The accuracy of ANN was higher in both the diabetes and cardiovascular disease cases (87.29 in diabetic patients and 89.38 in cardiovascular disease patients, respectively) when compared to the results of a set of 10 scientific articles on diabetes classification and 10 articles on cardiovascular disease. The Nave Bayesian network utilised may not be as reliable as the ANN method due to the observed nodes' independence. As a consequence, using an artificial neural network for diabetes categorization has been shown to improve accuracy with respect to NCL and/or CVD.

Clinical decision support systems in the past have often been constructed on the basis of either a single classification model or a generic set of these models with middling performance. In an effort to aid in the accurate diagnosis of cardiovascular disease (CVD), Eom et al. [6] have developed a classification-based strategy. The AptacidSS-E system overcomes the shortcomings of conventional methods by employing a number of different classifiers. According to recent polls, heart disease is one of the leading killers, yet early detection and treatment can save many lives. They have come up with a way to diagnose CVD by putting together the results of four different classifiers into a single output.

Neural networks and support vector machines both have a solid reputation as first-rate classifiers. As an added bonus, decision trees and Bayesian networks have been developed to further boost the system's efficacy. Four aptamer-based biochip datasets were used for training and testing the system, with CVD data from 66 samples being used as one example. Data inaccuracies are reduced with the use of three additional datasets. By comparing their results to those of models based on a single way to group things, the authors have looked at the benefits of a complex system with multiple ways to group things.

A cross-qualification test was used to assess the AptaCDSS-E system's potential to provide the promised results. The authors say that the experimental results of their system show that it can be used to make clinical diagnosis decisions because it has high diagnostic accuracy (> 94%) and small expected variation intervals (6%).

There has been a worldwide increase in interest from scientists interested in using medical data sets. Medical data sets for assessing illness groups are a common use of data mining techniques. By utilising clustering, noise reduction, and evaluation techniques, Nilashi et al. [7] offer a unique knowledge-based approach for illness assessment. As a means of articulating fuzzy rules inside a knowledge-based system, the authors have turned to classification and regression trees (CART). Multiple tools found in public health data sets have been tried and tested by the authors. The Indian Diabetes Pima, Mesothelioma, Wisconsin Diagnostic Breast Cancer, Cleveland, Statlog, and Parkinson Telemonitoring Datasets show that the suggested technique may greatly enhance illness prediction accuracy. Predicting illness from

large, real-world medical datasets using fuzzy rules, CART with noise reduction, and grouping approaches shows promising results. Knowledge-based systems, which do clinical analysis, aid doctors in their care.

Multiple tests were run using publicly available medical data sets to evaluate the efficacy of the suggested strategy and validate the system. The Indian Diabetes Pima, Mesothelioma, Wisconsin Diagnostic Breast Cancer, Statlog, Cleveland, and Parkinson Telemonitoring datasets all hail from the University of California, Data Mining Repository, Irvine (UCI). The results show that a combination of the clustering approach, principal component analysis, and fuzzy rule-based methods improves the accuracy of forecasts.

This paper proposes an evaluation method for using public UCI datasets with input and output parameters for a targeted diagnostic. More importantly, unlike the big data on healthcare, the data in these sets is not particularly complicated. Not only that, but large amounts of health care data

Because of progress in AI, many important parts of human life have changed in big ways.Machine learning (ML), a branch of artificial intelligence that automates the retrieval of data from vast databases, is finding significant use in the field of cardiovascular medicine. An introduction to the ML techniques employed in building inferential and data-driven models has been presented by Al'Aref et al. [8]. Echocardiography, electrocardiography, and newly emerging non-invasive imaging methods such as coronary artery calcium measurement and coronary CT angiography are only a few of the areas of applicability in ML that have been emphasised. They talked about what they had found by pushing the limits of how machine learning (ML) algorithms are used in cardiology right now. Rapid digitalization in healthcare presents an opportunity to use ML to find answers to critical medical problems. Traditional statistical approaches are still widely used in the medical research community, but ML has offered a new toolset to deal with the industry's fast evolution. ML can also provide a strong platform for combining clinical data with imaging data, which is helpful for a wide range of complex cardiac conditions like heart failure.

Recent uses of ML in cardiology, particularly in cardiac imaging, are discussed in this study. Machine learning could change the way medical research is done by optimising the clinical workflow in a way that reduces risk and improves effectiveness.

Previous research has found a link between prenatal exposure to air pollution and an increased risk of CHD. However, the impact of aerodynamic diameter particles of 10 microns (PM10) on CHD is uncertain. The non-linear features of PM10 exposure on the risk of coronary heart disease throughout the crucial period, i.e., weeks 3–8, were investigated by Ren Zhoupeng et al. [9] using two machine learning algorithms, namely Gradient Boost (GB) and Random Forest (RF). A cohort analysis of 39,053 live births in Beijing was done by the authors between 2009 and 2012. The coronary heart disease coefficient related to prenatal PM10 exposure after adjusting for maternal and perinatal variables was studied using GB and RF methods. Machine-learning algorithms agree that exposure to PM10 during pregnancy is a major risk factor for CHD.

PM10 appears to be more vulnerable to coronary artery disease between weeks 3 and 8 of pregnancy, which is within the prenatal heart development window. The complicated and linear association between maternal exposure to air pollution and the risk of birth abnormalities may be investigated with the use of machine learning algorithms. Studies are needed to figure out how important different times and types of air pollution are when it comes to the risk of coronary heart disease.

Classification is an effective and widely used machine learning assessment method. The accuracy of forecasts made by certain classifiers is good enough, while those made by others are severely lacking. Ensemble classification, as mentioned by Latha et al.[10], is a technique for boosting the performance of inefficient

algorithms by integrating several kinds of classifications. Testing of this instrument was conducted on a dataset including information about cardiovascular diseases.

A comparative analytical approach was used to see how generic technology can be used to improve the accuracy of heart disease predictions. The purpose of this article is not only to increase the accuracy of weak classification algorithms, but also to implement the algorithm with a medical dataset to demonstrate its effectiveness to predict the disease at an early phase. Indication from the results of the study of the general methods, such as packing and stimulation, are effective in improving the accuracy of predicting weak classifiers and have satisfactory results in determining the risk of heart disease. Advantage of the collective classification is that the weak classifiers have a maximum accuracy of 7%. Process performance has been further improved with the introduction of feature selection, and the results show a significant improvement in forecasting Accuracy.

III. PROPOSED SYSTEM

This paper presents an analysis of several machine learning techniques for predicting cardiovascular disease based on patient medical records.In this paper, we will first collect the necessary data, then perform some preliminary processing on it (for example, filling in blanks with the average, median, or mode of the entire attribute), and finally, split the resulting data in half. On the training dataset, we train the machine learning algorithms to produce a test model; in this case, we use KNN (K-Nearest Neighbors), Naive Bayes, Support Vector Machine, Logistic Regression, ADABoost, and XGBoost as voting classifiers. The testing dataset contains only 20% of the data. Below is the flow chart for the proposed system approach as shown in figure 1.:

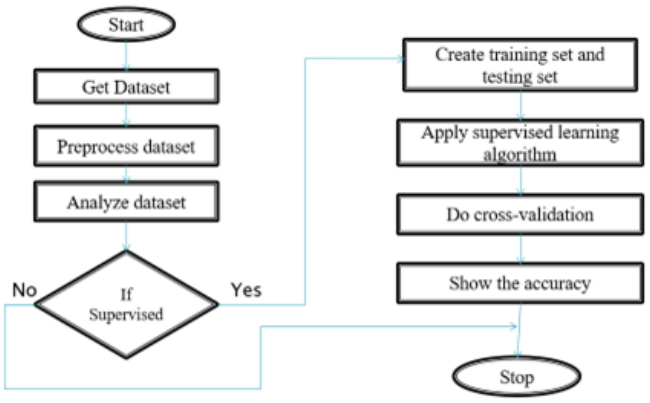


Fig1: Flow chart of the proposed work

Initially, we feed in the heart failure data collection that has been gathered. Many types of machine learning algorithms are then applied to the data. If it's a supervised learning algorithm, we'll set up training and testing set and use that to refine our model. After that, you do something called "cross-validation" to make sure everything went smoothly.

After the cross-validation procedure has been completed successfully, the system will display the algorithm's accuracy. If it turns out that the dataset isn't suitable for a supervised learning algorithm, the system will immediately stop and move on to the next one.

The below figure illustrates the various strategies used in prediction of heart disease.

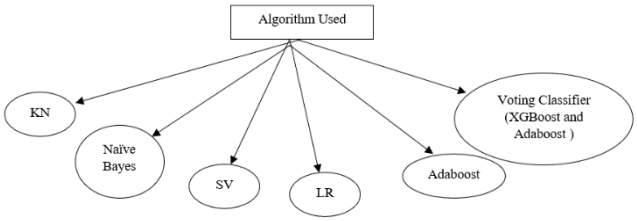


Fig2: Used Machine Learning Techniques

The formula may be utilised in forecasting. For this research, we used a number of different machine learning techniques, including the KNN, NB, SVM, Logistic Regression, ADABoost, and XGBoost voting classifiers.

K-Nearest Neighbours (KNN)

The KNN algorithm is used in non-parametric machine learning. The KNN algorithm is an example of supervised learning. This means the algorithm is

making a complete data call and using that information to make a prediction. However, the deviation from the Euclidean distance is the standard measure.

Support Vector Machine (SVM)

By presenting the training data set on a hyperplane that separates the points into two groups, the SVM method illustrates the presence and absence of heart disease. Hyper planes that maximise the distance between two dimensions are central to SVM's operation. Class variance is a problem in machine learning when there is an imbalance between the number of positive and negative features. The classifier will be ineffective unless this distinction is taken into account.

Naive Bayes Algorithm (NB)

This is a classification method used when the input dimensionality is very large. According to the Naive Bayes classifier, the presence of one feature in a class is unrelated to the presence of another feature. It is backed up by the Bayes theorem.

Logistic Regression Algorithm (LR)

Logistic regression (LR) models are trained with five splitting conditions and tested with test data for prediction to obtain the highest accuracy and to discover the models' behaviour, which is then used to make predictions about cardiac disease using the ML model. The algorithm's 1 or 0 result categories show whether or not a person has heart disease.

Adaboost Algorithm

In the annals of machine learning, AdaBoost (Adaptive Boosting) stands out as the first boosting algorithm to successfully combine multiple weak classifiers into a single robust one. Classification problems, such as binary classification, are the primary focus.

Multiple medical centres with cardiac patients and healthy controls have contributed to the data. Typical human information includes the 14 characteristics listed in Table 1.

There are approximately 1024 values in the dataset, which will be split into two parts. 80% of the data is used to train the model, and 20% is used to test the model.

Dataset used: **Number of Instances:** 20000

Number of Attributes: 14

For Each Attribute: (all numeric-valued)

For Each Attribute: (all numeric-valued)

age	Sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target

IV. Results and Analysis

The parameters used for performing the machine learning process across the considered algorithms are tabulated below.

Table 1.

S.No.	Name of the Machine Learning Algorithm	Parameters considered and its values
1	K-Nearest Neighbors	**n_neighbors=5, n_jobs=-1, leaf_size=60, algorithm=brute**
2	Gaussian Naïve Bayes	**No parameters are considered for learning**
3	Support Vector Machine	**kernel=linear**
4	Logistic Regression	**random_state = 0**
5	Adaboost	**n_estimators=100, random_state=0**
6	Voting Classifier (XGBoost and AdaBoost)	**AdaBoostClassifier(n_estimators=100,**

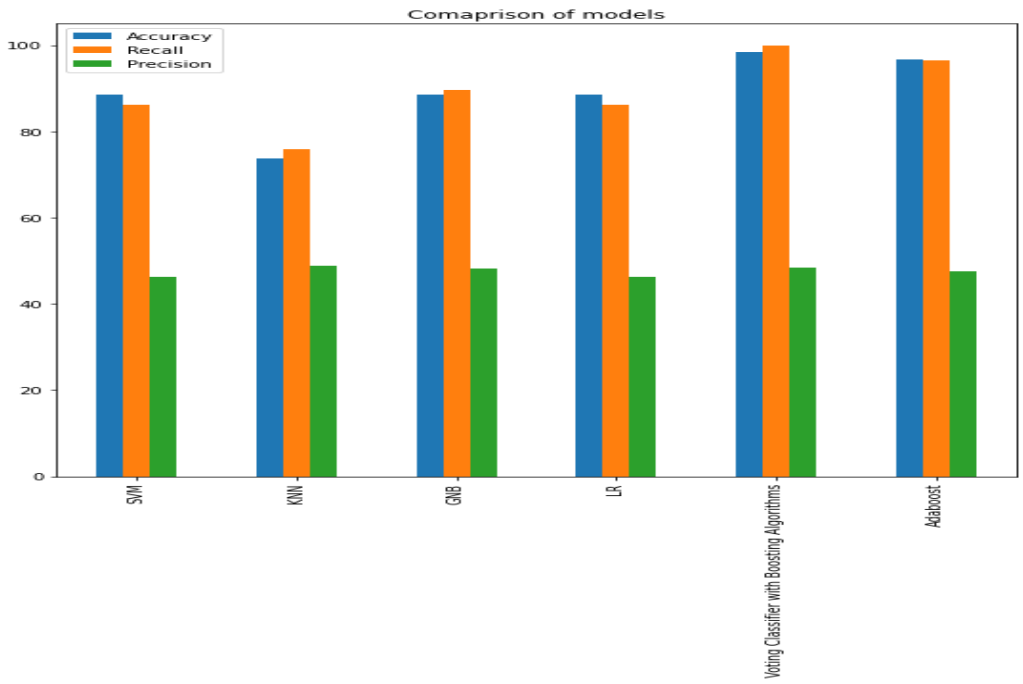
		random_state=0) **XGBClassifier(objective="binary:logistic", random_state=42)** **VotingClassifier(estimators = estimator, voting ='hard')**

The performance of the considered machine learning models in terms of findings for accuracy is tabulated as below.

Table 2 : Prediction Accuracy Results

Method	Accuracy
KNN	73.77%
NB	88.52%
SVM	88.52%
Logistic Regression	88%
Adaboost	96%
Voting Classifier(XG Boost+Adaboost)	98.3%

From the above table 2. It is observed that Voting Classifier i.e., XGBoost and AdaBoost method has the best accuracy. The Voting Classifier technique produced the most accurate findings. This is because the XGBoost algorithm learns the patterns in the data in parallelized manner and runs faster than other gradient boosting implementations. The AdaBoost algorithm makes use of multiple decision stumps and is less prone to overfitting. The experimental results are visualized in the figure provided below.



The confusion matrices of the implemented KNN and SVM algorithms are tabulated below.

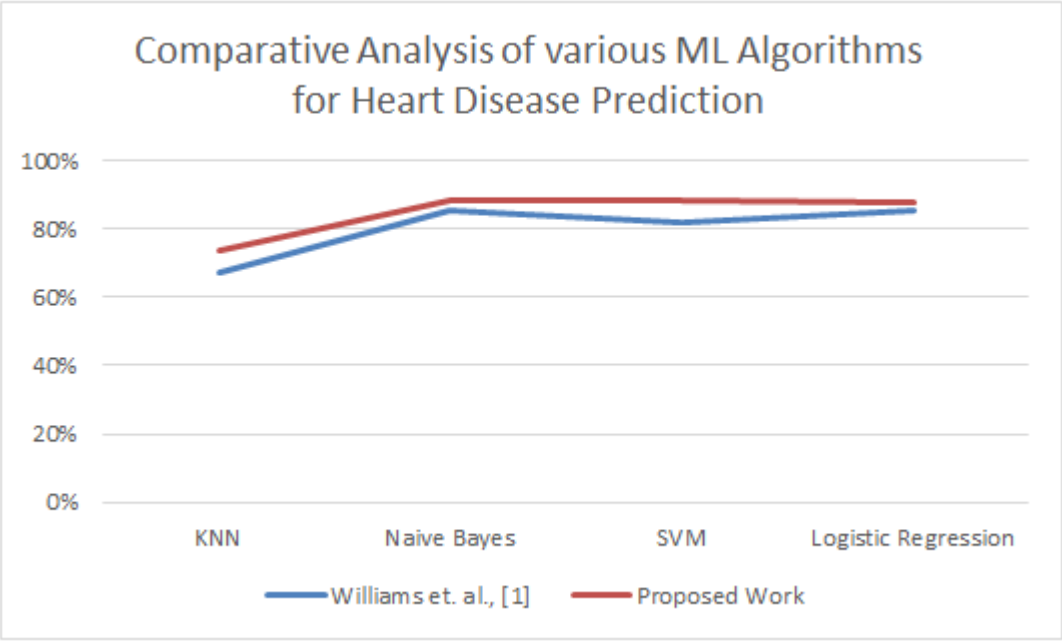
ML Algorithm	Confusion Matrix
KNN	array([[22, 7], [9, 23]], dtype=int64), Number of records in test set = 61
SVM	array([[25, 4], [3, 29]], dtype=int64), Number of records in test set = 61

It is observed from the above table that the Type-I and Type-II error counts on 20% of the test data are very less. This helps to draw the important inference of the very good performance of the KNN and SVM algorithms. When compared in between KNN and SVM, SVM classifier has performed better than KNN classifier in terms of reduced Type-I and Type-II error counts.

Table: 3. COMPARATIVE ANALYSIS:

Author and year	Title	Techniques	Accuracy
Reldean Williams, TokozaniShongwe, Ali N. Hasan, and Vikash Rameshar,2021 **[1]**	Heart Disease Prediction using Machine Learning Techniques	KNN, Naive Bayes, SVM, Logistic Regression	67.21% 85.25% 81.97% 85.25%
Proposed Work	Prediction of Heart Disease using Machine Learning Algorithms	KNN, Naive Bayes, SVM, Logistic Regression, **Adaboost,** **Voting Classifier**	73.77% 88.52% 88.52% 88% **96%** **98.3%**

The comparison of the proposed work with the work of [1] is presented in the below figure.



It is observed from the above figure that across all the machine learning algorithms that are compared there is an average increase of 4.78% accuracy in terms of performance of all the algorithms. This specifies that the considered parameters for learning the patterns across the machine learning algorithms have improved heart disease prediction accuracies in the individual manner and on an average.

The effectiveness of the algorithms is evaluated by the machine learning model. The results are summarised below in terms of precision:

V. CONCLUSION

KNN, Naive Bayes, Support Vector Machine, Logistic Regression, Adaboost, and Voting Classifier (XGBoost + Adaboost) are just some of the machine learning techniques that can be used to make predictions about cardiovascular disease. According to the findings, the voting classifier performed better than the other methods and attained 98% prediction accuracy. Using ensemble methods and artificial neural networks to learn more about the dataset could also be helpful.

VI. REFERENCES

[1]. R. Williams, T. Shongwe, A. N. Hasan and V. Rameshar, "Heart Disease Prediction using Machine Learning Techniques," 2021 International Conference on Data Analytics for Business and Industry (ICDABI), 2021, pp. 118-123, doi: 10.1109/ICDABI53623.2021.9655783.

[2]. S. R. Tithi, A. Aktar, F. Aleem and A. Chakrabarty, "ECG data analysis and heart disease prediction using machine learning algorithms," 2019 IEEE Region 10 Symposium (TENSYMP), 2019, pp. 819-824, doi: 10.1109/TENSYMP46218.2019.8971374.

[3]. Purushottam, Kanak Saxena, and Richa Sharma. "Efficient Heart Disease Prediction System Using Decision Tree." International Conference on Computing, Communication & Automation, 2015.https://doi.org/10.1109/ccaa.2015.7148346.

[4]. Li, Yan, Matthew Sperrin, Glen P. Martin, Darren M. Ashcroft, and Tjeerd Pieter Van Staa. "Examining the Impact of Data Quality and Completeness of Electronic Health Records on Predictions of Patients' Risks of Cardiovascular Disease." International Journal of Medical Informatics 133 (2020): 104033. https://doi.org/10.1016/j.ijmedinf.2019.104033.

[5]. Alic, Berina, Lejla Gurbeta, and Almir Badnjevic. "Machine Learning Techniques for Classification of Diabetes and Cardiovascular Diseases." 2017 6th Mediterranean Conference on Embedded Computing (MECO), 2017.https://doi.org/10.1109/meco.2017.7977152.

[6]. Eom, J, S Kim, and B Zhang. "AptaCDSS-E: A Classifier Ensemble-Based Clinical Decision Support System for Cardiovascular Disease Level Prediction." Expert Systems with Applications 34, no. 4 (2008): 2465–79. https://doi.org/10.1016/j.eswa.2007.04.015.

[7]. Nilashi, Mehrbakhsh, Othman Bin Ibrahim, Hossein Ahmadi, and Leila Shahmoradi. "An Analytical Method for Diseases Prediction Using Machine Learning Techniques." Computers & Chemical Engineering 106 (2017): 212–23. https://doi.org/10.1016/j.compchemeng.2017.06.011.

[8]. Al'Aref, Subhi J, Khalil Anchouche, Gurpreet Singh, Piotr J Slomka, Kranthi K Kolli, Amit Kumar, Mohit Pandey, et al. "Clinical Applications of Machine Learning in Cardiovascular Disease and Its Relevance to Cardiac Imaging." European Heart Journal 40, no. 24 (2018): 1975–86. https://doi.org/10.1093/eurheartj/ehy404.

[9]. Ren, Zhoupeng, Jun Zhu, Yanfang Gao, Qian Yin, Maogui Hu, Li Dai, Changfei Deng, et al.

"Maternal Exposure to Ambient PM10 during Pregnancy Increases the Risk of Congenital Heart Defects: Evidence from Machine Learning Models." Science of The Total Environment 630 (2018): 1–10. https://doi.org/10.1016/j.scitotenv.2018.02.181.

[10]. Latha, C. Beulah Christalin, and S. Carolin Jeeva. "Improving the Accuracy of Prediction of Heart Disease Risk Based on Ensemble Classification Techniques." Informatics in Medicine Unlocked 16 (2019): 100203.

Cite this article as :

Yegamati Akhila, R. Usha Rani, "Prediction of Heart Disease Using Machine Learning Algorithms", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 5, pp. 273-282, September-October 2022.
Journal URL : https://ijsrcseit.com/CSEIT228551